Running head: REFERENTIAL LOOKING PREDICTS WORD LEARNING

Attention to conversational referents predicts word learning in young children

Michael C. Frank

Department of Psychology, Stanford University

Allison M. Kraus

Department of Psychology, Stanford University

Theresa Hennings

Department of Psychology, University of Washington

Abstract

Language learners face two challenges in parallel: interpreting the speech they hear and figuring out what individual words mean. Recent models of word learning have suggested that these two processes are coupled, such that online disambiguation interacts with longer-term cross-situational learning. We provide a direct test of this hypothesis, using eye-tracking to monitor the fixations of a large, diverse sample of children from 1 – 5 years of age as they watched a series of naturalistic videos that taught novel words. Although older children learned word-object mappings better, these developmental effects were completely mediated by in-the-moment attention to the appropriate object. Our results support a view of word learning as driven by communicative inferences that guide attention and provide inputs to processes of learning and memory.

At dinner, a young child sits in the highchair as her father alternates feeding her carrots and peas. He lifts a spoonful of carrots and says “Do you like the carrots?” A few bites later, she grimaces and rejects the spoon. He says “Do you want some peas instead?” From the child’s perspective, linguistic situations like this one can be very ambiguous. Does the word “carrots” mean carrots, does it mean peas, or does it have an entirely different meaning?

Word learners make a wide variety of inferences to overcome the natural ambiguity of communicative situations like this one. For example, social cues—like the parent pointing at the peas—could provide a clear signal to the child of the speaker’s intended referent (Baldwin, 1993; Hollich, Hirsh-Pasek, & Golinkoff, 2000). If she has already learned some food vocabulary, the child could also use her knowledge of the word “carrots” to help her infer what peas are (Markman & Wachtel, 1988; Clark, 1988). Or she could gather partial knowledge from this learning situation—that peas and carrots both co-occur with the word “carrots”—and another dinner, perhaps one with carrots and bananas, could strengthen her cross-situational association between the word and the correct food (Yu & Smith, 2007; L. Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009).

These inferences differ not only in the kind of information they require, but also in the timescale over which they operate. Social cues like pointing or eye-gaze indicate likely referents of the current utterance and can be used in the moment, without prior context. Inferences based on linguistic knowledge can also be applied to a single utterance. In contrast, cross-situational word learning involves the gradual accumulation of information across many different exposures to a word, separated in time.

Yet recent computational work has suggested the hypothesis that rather than being distinct strategies, these varied ways of gaining insight into word meaning are different facets of the same unified process. Two recent models posit that word learning is a process of inference at two timescales: referent selection in the moment and long-term mapping (Frank, Goodman, & Tenenbaum, 2009; McMurray, Horst, & Samuelson, 2012). While on a “pure” cross-situational account, all words and objects in a context are associated with one another, on a two-timescale account, the learner reaches an interpretation of which referent is intended. This interpretation then provides the input for long-term word learning (for a lower-level, attentional account, see Yu & Smith, 2012). In the language of the example above, if the child is able to converge on the carrots as the referent—whether by lexical, social, or pragmatic means—then carrots (and critically, not peas) are the appropriate object for mapping.

Support for this hypothesis comes from artificial language experiments where multiple words and objects are presented on each trial. Infants in such experiments can learn word-object mappings from individually-ambiguous presentations (L. Smith & Yu, 2008; Yu & Smith, 2010; L. Smith & Yu, in press). In one study, those infants who shifted between referents less and showed greater predictability in their eye-movements during training were more successful at identifying correct pairings at test (Yu & Smith, 2010); in another, those learners who best overcame their preference for novel stimulus items learned mappings better (L. Smith & Yu, in press). In these studies, both learners' accumulating knowledge and their attentional abilities can work together to guide their fixations to appropriate targets, gating their evidence about word-object mappings.

In artificial language experiments, words are typically heard in isolation, without a speaker, a pragmatic goal, or a linguistic context, however, and these basic communicative elements may alter the learning task dramatically. Learners' distribution of attention is especially likely to be affected by the presence of speakers, whose faces and hands provide powerful attentional signals even in complex and naturalistic displays (Frank, Goodman, Tenenbaum, & Fernald, 2009; Frank, Vul, & Saxe, 2012; Gliga, Elsabbagh, Andravizou, & Johnson, 2009). Hence, it is unknown whether in-the-moment attention to conversational referents (rather than, e.g., attention to the speakers producing the labels) is an adaptive strategy for word learning in more natural environments. Our current study was designed to address this question.

We created a video stimulus that taught children two novel words through a series of object-focused dialogues and monologues ("word learning videos"). These videos were intended to

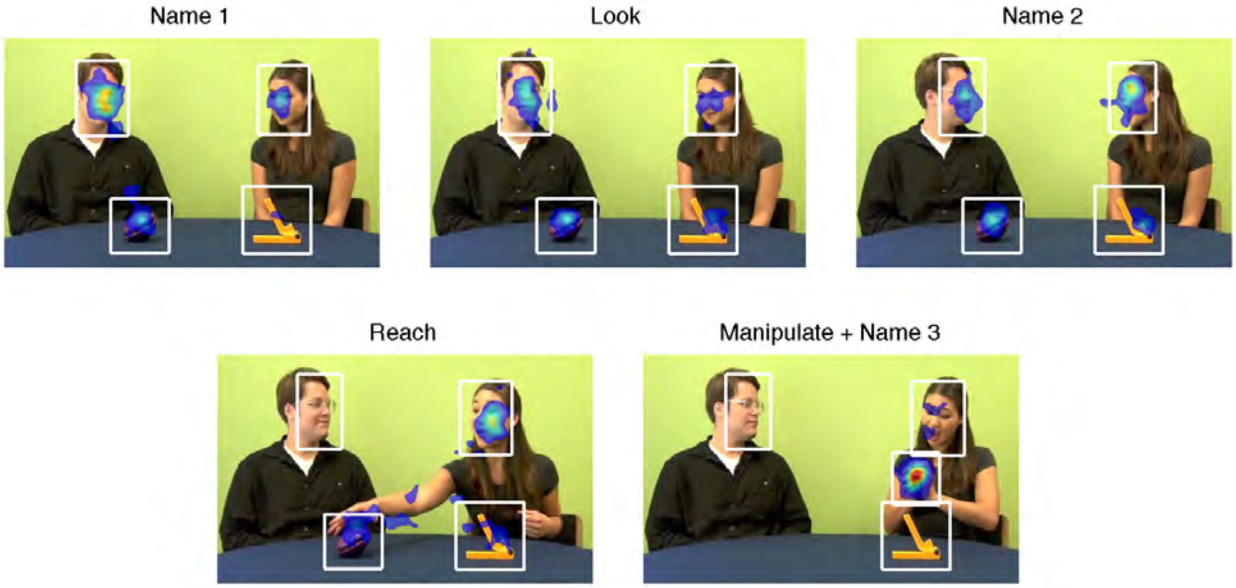

Figure 1. Example frames from the first word learning dialogue in our video stimulus. Each image shows the regions of interest used for later analysis (white boxes) and a heat map of the entire participant group's average point of gaze (hotter colors indicate more fixation; scale is constant across frames).

be sufficiently difficult in their structure that in-the-moment disambiguation would pose a significant challenge to young learners. We measured children's eye-movements as an index of their online inferences about the current conversational referent, and then tested their retention of the words they learned via a series of forced-choice test trials. These rich time-course data allowed us to test the hypothesis that those children who were successful in attending to the conversational referent would be the same children who learned and retained the words.

Methods

Participants

Families at the San Jose Children's Discovery Museum were invited to participate. We collected both demographic and eye-tracking data from 343 children in the target age range (1 – 5 years), of whom 113 were excluded from the final sample for at least one of the following reasons:

developmental issues (N=7), experimental issues including sibling interference (N=17), reported household English usage less than 75% (N=44), or unacceptable calibration (N=58). Our final sample included 229 children, ages 1 – 5 (M=3.2 years; age 1 – 2: N=35, age 2–3: N=60, age 3–4: N=72, age 4–5: N=62; 115 girls).

Stimulus and Procedure

We constructed an engaging video with three sections: calibration, learning, and test. Calibration began with a short video of a puppet to capture children's attention, and then showed an image of the puppet moving around to two calibration points. The learning section was a video montage (available at `http://langcog.stanford.edu/materials/reflook.html`), consisting of three distinct stimulus types: four word learning clips, two short distractor clips containing rich social interactions but no language (the "multiple people" videos of Frank et al., 2012), and two calibration check stimuli (a precessing annulus).

Word learning clips consisted of two monologues and two dialogues. Each clip contained either two novel objects (dialogues, Figure 1) or one novel object and one familiar object (monologues). Actors described each object. Each description sequence contained a first naming phrase, a look at the object accompanied by a comment, a second naming, a reach for the object, and then a demonstration of the object's function accompanied by a third naming. Each novel object appeared in three videos (two dialogues, one monologue) and was named nine times in total over the course of the video.

The test section consisted of 16 two-alternative forced-choice trials between pairs of pictures (Fernald, Zangl, Portillo, & Marchman, 2008). Eight tested familiar objects (e.g. dog/car or lamp/carrot); the other eight paired the two novel objects. Naming phrases were of the form "Look at the [car/fep]! Do you see it?" and were spoken by the actors in the video. Test trials were intermixed with a small number of filler items (e.g. a picture of a train) and one further distractor video.

Families were escorted to a small room off the museum floor, where children watched the video from a car seat approximately 60cm from the monitor of an SMI RED 120 Hz corneal reflection eye-tracker (mounted on an adjustable arm). Total experiment duration was approximately 6 minutes.

Data Analysis

To ensure appropriate precision in region-of-interest analyses, infants' calibrations were corrected and verified via robust regression (described in Frank et al., 2012), and calibration corrections were assessed by two independent coders ($\kappa = .85$). For maximum accuracy in the analyses reported below, we excluded children whose calibrations could not be verified and corrected.

To assess looking behavior related to interpretation of the conversational referent, we created a measure of "referential looking": how much children looked at an object as it was being described in the word learning clips. We broke descriptions into six time periods (Figure 1): a baseline period of exactly 2 s, name 1 to look (M = 1.7 s), look to name 2 (M = 2.0 s), name 2 to initiation of reach (M= 4.8 s), initiation of reach to point of contact (M = .8 s), and after contact with the object (1 s). For each of these, we examined the proportion of total dwell-time that fell in a region-of-interest around the object.

To measure accuracy on test trials, we computed each child's total dwell time on the target following onset of the labeling word. For maximal comparability across age groups, we chose a 1500 ms window for the familiar words and a 2000 ms window for the novel words (both beginning 500 ms after word onset). The reliability of individual participants' measurements was related to the amount of data they contributed. We therefore created two exclusion criteria: a strict criterion—in which a participants' measurements for novel or familiar test trials were only included if they had 4/8 trials with $> 80\%$ data in each—and a permissive criterion—in which participants were included if they had 2/8 trials with $> 50\%$ data). Figures and results in the text use the strict inclusion criterion; qualitatively similar results from the permissive criterion are reported in Supplemental Information.

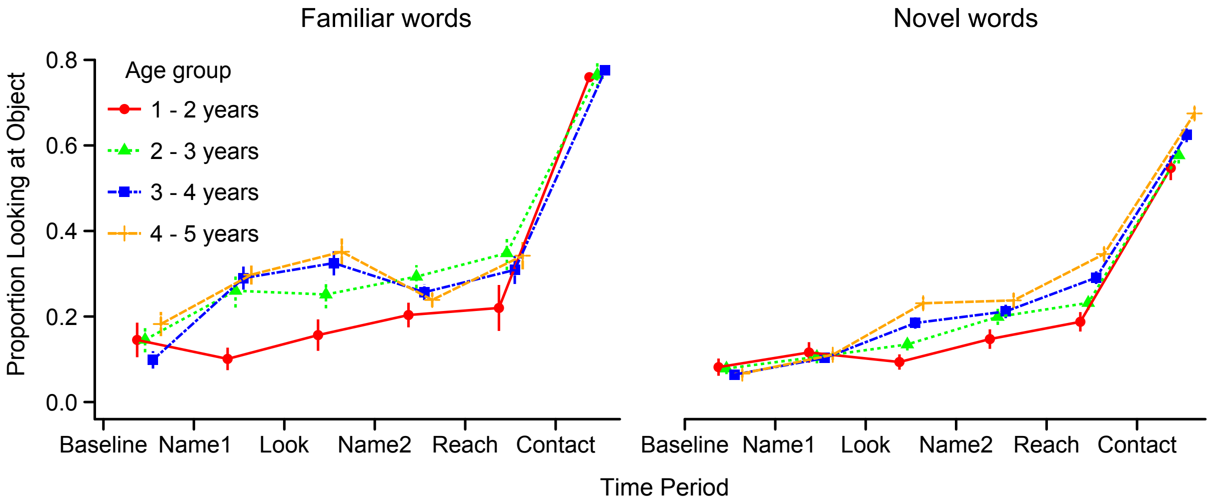

Figure 2. Proportion looking at the correct referent for familiar and novel words in the word learning videos. The horizontal axis shows markers in the description, so points falling between two markers summarize looking during that time period. Color and marker show different age groups; ranges show standard error of the mean. Points are slightly offset on the horizontal to avoid overplotting.

Results and Discussion

The primary question of interest for our study was whether we were able to predict accuracy at test based on looking behavior during the learning phase of the study. We began by examining learning and test phases separately and then explored the relationship between them.

During descriptions of the objects, looking to the target object gradually increased over time and was slightly higher for older children (Figure 2). We observed a large increase in looking to both novel and familiar objects once the speaker reached for the object: Before the point of contact, children's gaze was primarily on the faces of the speakers, while after the point of contact, children looked much more at the object that was being manipulated. The combination of a goal-directed action and a complex motion event in an otherwise static scene was enough to focus children's fixation across ages. In contrast, neither naming nor looking at objects consistently drove the majority of children to look at the speaker's intended referent.

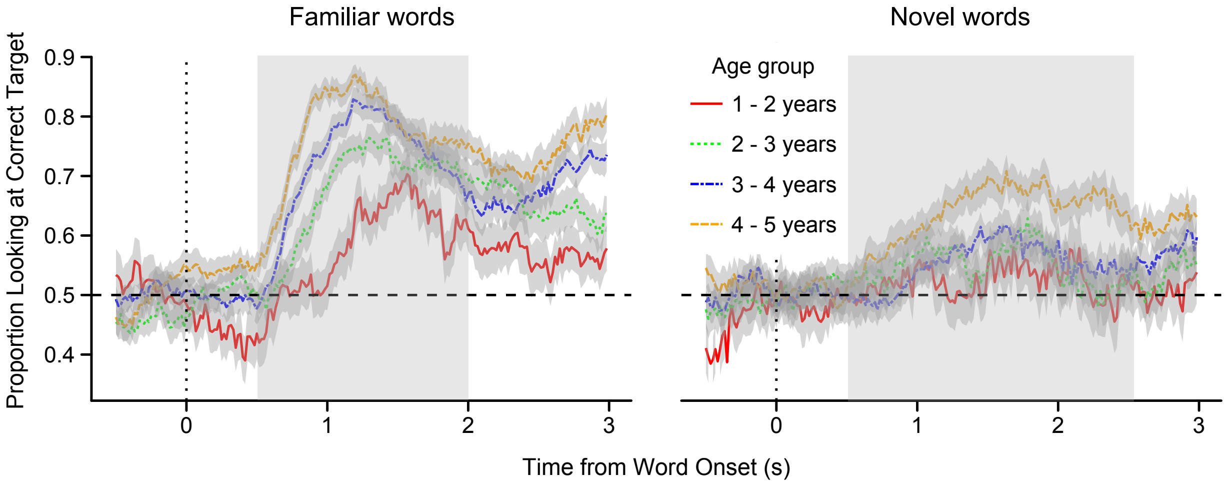

Figure 3. Time-course plots from familiar and novel word learning trials. Each line shows the proportion looking at the correct target, with the shaded area showing the standard error of the mean. The horizontal axis is numbered from the onset of the target word. Gray regions mark the window for further analyses.

Turning to the test trials, we found substantial developmental differences in both familiar and novel word recognition (Figure 3). Consistent with previous work, both children's speed of word recognition and their proportion looking to the correct target increased with age (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998; Fernald, Perfors, & Marchman, 2006). A linear model found a reliable relationship between age and familiar word recognition accuracy ($\beta = .056$, $p < .0001$, $r^2 = .15$), which persisted when controlling for gender and parent education (neither approaching significance as predictors within this relatively high-education sample). For novel words, performance was lower, but we found a similarly-sized effect of age ($\beta = .051$, $p = .0001$, $r^2 = .13$). Our stimulus allowed older children to learn novel words, but it was challenging overall for the youngest participants.

In the key test of our hypothesis, referential looking was a strong predictor of children's learning of object labels. We constructed a linear regression model which predicted novel word learning accuracy for each child as a function of referential looking during each of the target

Predictor	Estimate	Std. Error	t value	p value	
Baseline	0.18	0.18	1.10	0.27	
Name 1	0.04	0.14	0.29	0.77	
Look	0.37	0.13	2.93	0.004	**
Name 2	0.05	0.11	0.45	0.66	
Reach	0.21	0.12	1.73	0.09	.
Point of contact	0.12	0.10	1.18	0.24	

Table 1

Coefficients along with standard error and significance information for a simple least squares regression predicting novel word learning based on looking time to the object during the baseline, name 1, look, name 2, reach, and point-of-contact portions of the movie (intercept omitted). . indicates $p < .1$, ** indicates* $p < .05$, *and ** indicates* $p < .01$.

periods (coefficient estimates are reported in Table 1). Overall, the model was highly significant ($p < .0001$, $r^2 = .29$), and the strongest predictor of word learning was gaze following behavior immediately after the actors looked at the novel objects.

Neither total time attending to the video during the referential looking task ($F(1) = .91$, $p = .34$) nor familiar word recognition accuracy ($F(1) = .47$, $p = .49$) significantly improved model fit, ruling out confounds of attentiveness during training or test phases. Independent measures of preference to look at faces and hands in the dynamic distractor videos also did not increase fit ($F(1) = .03$, $p = .86$ and $F(1) = .69$, $p = .41$, respectively), ruling out basic social preferences as explanatory factors.

Finally, referential looking better predicted novel word learning than did infants' chronological age (Figure 4). The referential looking model completely mediated the effect of age (Baron & Kenny, 1986): Age was a significant predictor alone ($\beta = .051$, $p < .0001$), but when the referential looking composite was added to the regression model, the coefficient estimate for

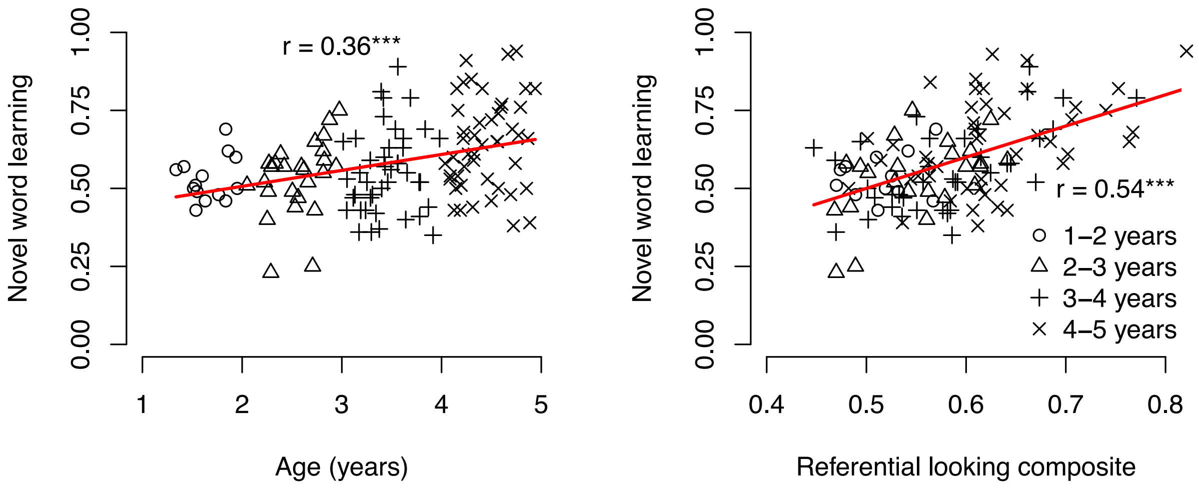

Figure 4. The relationship between novel word learning and age (left) and between novel word learning and referential looking (right). Markers show participants in different age groups; lines are best-fit lines from a simple linear regression. Significance values are marked with *** for $p < .001$.

age declined substantially and was no longer significant ($\beta = .012$, $p = .35$). This mediation result held for children under 3 years old, indicating that it was not driven by older children's performance.

General Discussion

Children in our study watched a series of short video clips that were intended to challenge their word learning skills. While older children with more language experience created stronger mappings between words and their corresponding objects, this developmental effect was mediated by their in-the-moment attention during the video. The older children's advantage in learning was a function of their ability to direct their attention to the appropriate referent. This finding provides support for models of word learning that posit a strong linkage between momentary ambiguity resolution and long-term learning (Frank, Goodman, & Tenenbaum, 2009; McMurray et al., 2012) and replicates in a single session previous work suggesting that social intention—especially the ability to follow gaze and gestures—is related to vocabulary growth (Brooks & Meltzoff, 2005, 2008; Carpenter, Nagell, & Tomasello, 1998).

Of course, there are many developmental factors involved in children's increasing proficiency as word learners, including faster language processing abilities (Fernald et al., 1998) and better understanding of social interactions (Wellman & Liu, 2004), as well as other more general changes in attentional and memory abilities (Gathercole, Pickering, Ambridge, & Wearing, 2004). Our data provide insight into one important route by which these factors affect vocabulary growth: by helping children monitor the topic of conversation.

Although we have interpreted our findings in terms of children's understanding of reference, a lower-level interpretation is possible as well. The key factor in learning word-object associations could be the child's attention to the correct location during naming—even absent any understanding of reference per se. We believe this kind of purely associative account is unlikely to be correct, however. In our data, it does not explain why fixation during naming was less predictive of mapping than fixation during looking or reaching. (An intentional account would suggest that these actions are predictive because they signal reference.) In addition, a pure associative account does not provide any interpretation for classic findings in which immediate visual attention is dissociated from learning (Baldwin, 1991, 1993). Nevertheless, future work could provide stronger causal evidence for this claim by directly manipulating the source of visual attention to objects.

In distinguishing "carrots" from "peas," referential uncertainty is one of the major challenges facing early word learners. Theories of word learning initially focused on the strategies by which reference can be determined within a single learning situation (Markman & Wachtel, 1988; Baldwin, 1993; Hollich et al., 2000), but the discovery that learners can infer mappings across situations has led to an explosion of work investigating the scope and limits of cross-situational learning as well (L. Smith & Yu, 2008; K. Smith, Smith, & Blythe, 2011; Trueswell, Medina, Hafri, & Gleitman, in press). Yet it is unlikely that we can fully understand one process in the absence of the other. Our work here adds to the growing body of evidence suggesting that the interplay between two aspects of learning—in the moment and across situations—is crucial for understanding children's word learning abilities.

References

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*, 874–890.

Baldwin, D. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, *20*, 395–395.

Baron, R., & Kenny, D. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173.

Brooks, R., & Meltzoff, A. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*, 535–543.

Brooks, R., & Meltzoff, A. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. *Journal of Child Language*, *35*, 207–220.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63(4)*.

Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, *15*, 317–335.

Fernald, A., Perfors, A., & Marchman, V. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*, 98.

Fernald, A., Pinto, J., Swingley, D., Weinberg, A., & McRoberts, G. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*, 228–231.

Fernald, A., Zangl, R., Portillo, A., & Marchman, V. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. Sekerina, E. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (Vol. 44, pp. 97–135). Amsterdam/Philadelphia: John Benjamins.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Frank, M. C., Goodman, N. D., Tenenbaum, J. B., & Fernald, A. (2009). Continuity of discourse provides information for word learning. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Frank, M. C., Vul, E., & Saxe, R. (2012). Measuring the development of social attention using free-viewing. *Infancy*, *17*, 355–375.

Gathercole, S., Pickering, S., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental psychology*, *40*, 177.

Gliga, T., Elsabbagh, M., Andravizou, A., & Johnson, M. (2009). Faces attract infants' attention in complex displays. *Infancy*, *14*, 550–562.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831.

Smith, K., Smith, A., & Blythe, R. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*, 480–498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Smith, L., & Yu, C. (in press). Visual attention is not enough. *Language, Learning, and Development*.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (in press). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729–742.

Vouloumanos, A., & Werker, J. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, *45*, 1611–1617.

Wellman, H., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*, 523–541.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Yu, C., & Smith, L. (2010). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*, 165–180.

Yu, C., & Smith, L. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological Review*, *119*, 21.